Husseini Mohamad Ali

Rapport
- Projet Monte-Carlo et Chaînes de Markov -

SUJET : *“Google PageRank”*

Table des matières :

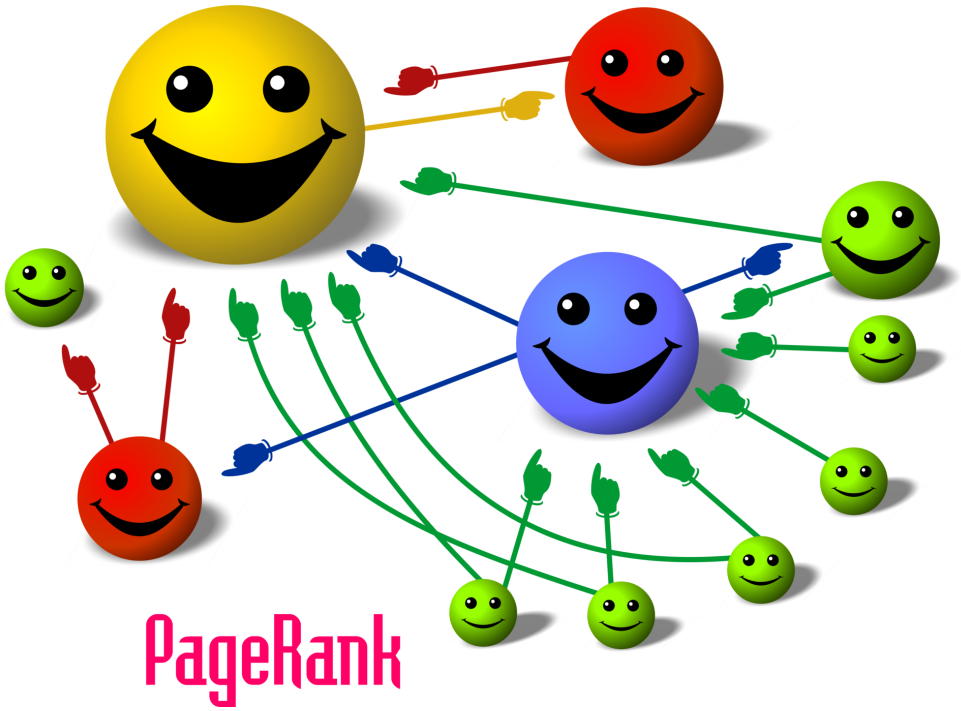
I) Introduction

L'algorithme PageRank, développé par Larry Page et Sergey Brin, repose sur les chaînes de Markov pour modéliser la navigation aléatoire sur le web et classer les pages de manière innovante. Avant cette méthode, les approches traditionnelles se limitaient à compter les liens entrants d'une page, sans évaluer la qualité des sites émetteurs. Cela posait problème : un lien provenant d'un site peu fiable ne devait pas avoir le même poids qu'un lien issu d'un site reconnu.

Page et Brin ont introduit le concept du "surfeur aléatoire", qui simule un utilisateur passant d'une page à l'autre via des liens ou effectuant des sauts aléatoires. Cette dynamique s'appuie sur une chaîne de Markov ergodique, garantissant une distribution stationnaire qui mesure la pertinence de chaque page.

Ainsi, le PageRank ne se contente pas de quantifier le nombre de liens, mais tient compte de la crédibilité des pages émettrices, permettant d'identifier les sites les plus influents dans un réseau complexe. Cette avancée a significativement amélioré la qualité des résultats proposés par les moteurs de recherche.

Dans la suite de ce rapport, nous présenterons d'abord les méthodes naïves et les difficultés qu'elles rencontrent (II), avant de détailler l'algorithme PageRank en lui-même (III) et de montrer pourquoi et comment il converge (III). Nous examinerons ensuite les attaques Sybil, le Google Bombing, tactiques destinées à manipuler les résultats (IV). Enfin, nous conclurons sur la place toujours centrale qu'occupe cet algorithme dans le classement des pages web (V).



II) Les méthodes naïves de classement et leurs limites

Avant que PageRank ne voit le jour, le critère dominant pour évaluer la popularité d'une page web reposait sur une approche des plus simples : compter ses liens entrants. En théorie, plus une page recevait de "votes" sous forme de liens, plus elle paraissait incontournable. Bien qu'intuitive à petite échelle, cette méthode a rapidement montré ses limites face à la complexité croissante du Web.

a. La fausse évidence des liens entrants

Compter uniquement le nombre de liens entrant peut sembler une mesure intuitive de sa popularité. Cependant, cette approche naïve présente des limites majeures. D'abord, elle repose sur une absence totale de pondération, traitant tous les liens de manière égale, quelle que soit leur provenance. Ainsi, un lien provenant d'un site influent est considéré avec le même poids qu'un lien issu d'un blog obscur ou de "fermes de liens" artificielles, ce qui fausse la perception de la qualité.

Ensuite, cette méthode est vulnérable aux manipulations. Certains acteurs mal intentionnés exploitent cette faiblesse en créant des pages factices, interconnectées, pour gonfler artificiellement le score de leurs sites cibles. Ces pratiques aboutissent à une surévaluation trompeuse de certaines pages, sapant la fiabilité du classement.

Par exemple, imaginons un réseau simplifié composé de trois sites :

1. *Site A*, un portail universitaire respecté, qui publie des recherches de pointe.
2. *Site B*, un blog personnel obscur, peu visité.
3. *Site C*, une ferme de liens artificielle, créée pour manipuler les classements.

Si l'on applique une méthode naïve, ces trois sites sont évalués uniquement en fonction du nombre de liens entrants qu'ils reçoivent. Voici ce qui se passe :

- *Site B* reçoit 50 liens provenant de *Site C*, alors que *Site A* n'en reçoit que 10, mais depuis d'autres portails universitaires fiables.
- Selon cette méthode, *Site B* sera considéré comme plus populaire que *Site A*, car il possède davantage de liens entrants, bien que la qualité des liens émis par *Site C* soit insignifiante et purement manipulatrice.

Cette surévaluation de *Site B* montre comment une méthode non pondérée est aveugle à la crédibilité des émetteurs de liens, dévalorisant des sites influents comme *Site A*.

b. Non-ergodicité et absence de téléportation.

Outre les faiblesses liées aux liens entrants, un autre problème fondamental apparaît lorsque l'on modélise la navigation avec un surfeur aléatoire : la non-ergodicité.

Certaines pages, appelées "sinks", n'ont aucun lien sortant. Une fois que le surfeur atteint l'une de ces pages, il se retrouve bloqué, incapable de poursuivre son exploration. D'autres parties du graphe fonctionnent comme des sous-graphes isolés : des groupes de pages qui ne sont connectées qu'entre elles. Les liens ne mènent jamais à l'extérieur de ce groupe ni depuis l'extérieur vers ce groupe. Autrement dit, le surfeur reste coincé à naviguer uniquement au sein de ce petit réseau sans jamais pouvoir explorer le reste du Web.

Ce problème est aggravé par l'absence de téléportation. Sans un mécanisme permettant au surfeur de "sauter" vers une page arbitraire, la chaîne de Markov associée n'est pas irréductible, ce qui empêche l'existence d'une distribution stationnaire globale. Le classement perd alors tout son sens, car il ne peut pas refléter une vision cohérente et complète du réseau.

Prenons maintenant un réseau où :

1. *Site D* contient uniquement des articles techniques et n'a aucun lien sortant.
2. *Sites E*, *F* et *G* forment un sous-graphe isolé, uniquement relié entre eux.
3. *Site H* est une page centrale avec des liens vers *D*, *E*, *F* et *G*, mais n'en reçoit aucun.

Si un surfeur aléatoire commence sur *Site H*, il peut accéder à *D*, mais reste bloqué une fois arrivé : il n'y a aucun lien sortant. De même, s'il atteint le sous-graphe (*E*, *F*, *G*), il ne pourra jamais en sortir, limitant drastiquement son exploration.

Sans un mécanisme de téléportation qui permettrait au surfeur de sauter vers une page aléatoire, la chaîne de Markov modélisant ce réseau devient non ergodique :

- Les "sinks" comme *D* interrompent la navigation.
- Les sous-graphes isolés faussent les résultats en concentrant le surfeur dans une portion réduite du réseau.

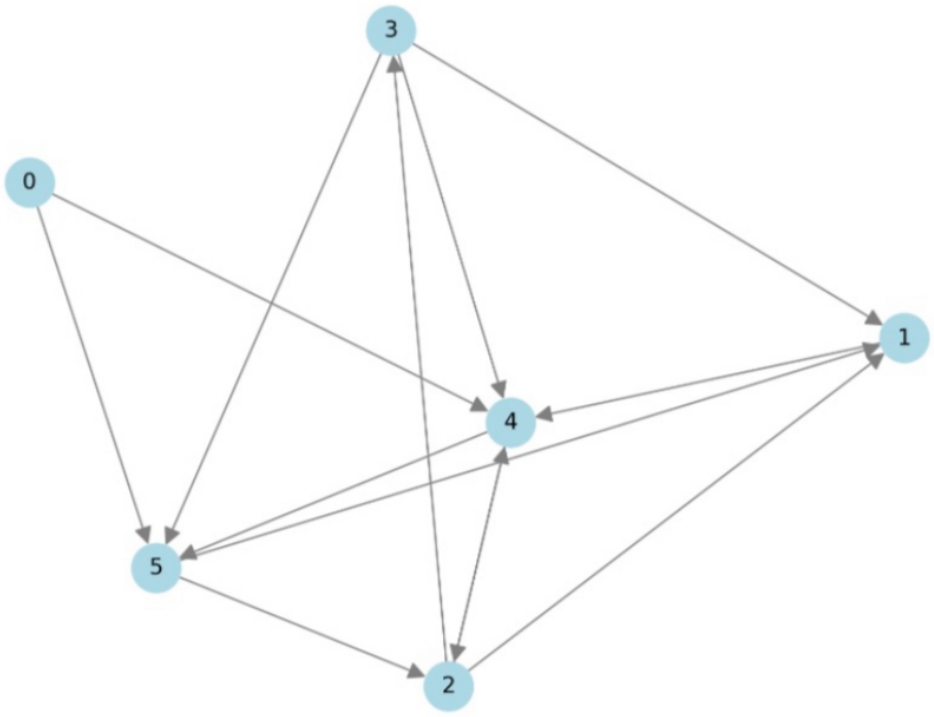
Ainsi, le classement naïf ne peut aboutir à une vision globale et cohérente du web. Ce genre de problème a motivé l'introduction du modèle de téléportation dans l'algorithme PageRank, garantissant une navigation fluide et un classement plus pertinent.

III) L'algorithme PageRank : principes & fonctionnement

Un comportement plus intéressant apparaît lorsque nous nous déplaçons aléatoirement sur un graphe orienté.

Dans ce graphe, toutes les arêtes ont un seul sens : les nœuds sont reliés non pas par des lignes, mais par des flèches. La chaîne peut se déplacer d'un sommet à un autre, mais uniquement dans les directions autorisées par les flèches.

Un exemple de graphe orienté est illustré par :



La matrice de transition pour ce graphe est : $P = \begin{bmatrix} 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}$

On peut conclure que cette matrice n'est pas régulière. Pourquoi ?

Une raison pour laquelle on peut affirmer cela est la présence d'une colonne de zéros (colonne 1). Toute puissance de P conservera cette colonne de zéros.

Pourquoi Page et Brin ont-ils choisi la marche aléatoire comme base de leur approche ? En réalité, l'intuition qui les a guidés était plus simple :

Leur idée était simplement de dire qu'une page est "importante" si de nombreuses pages "importantes" y sont liées.

Plus précisément, cette définition de l'"importance" est la suivante :

$$\text{L'importance de } k = \sum_j \left(\text{Importance de la page } j \cdot \text{Probabilité d'aller de } j \text{ à } k\right).$$

C'est une définition très intuitive de l'importance. Cependant, il y a un problème : L'importance de la page k est sur les 2 côtés de l'équation!

Comment résoudre cette équation pour obtenir une importance fixe pour une page donnée ? C'est ici que la marche aléatoire entre en jeu.

Ce que Page et Brin ont observé, c'est que cette équation :

$$\text{L'importance de } k = \sum_j \left(\text{Importance de la page } j \cdot \text{Probabilité d'aller de } j \text{ à } k\right)$$

est équivalente à : $x = xP$ si l'on encode :

- l'importance de toutes les pages dans le vecteur x , et
- la "probabilité de passer de la page j à la page k" dans la matrice stochastique P.

Nous sommes maintenant prêts à comprendre ce que Page et Brin disaient en 1998 :

« PageRank can be thought of as a model of user behavior. We assume there is a “random surfer” who is given a web page at random and keeps clicking on links, never hitting “back” but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. »

Le PageRank peut être considéré comme un modèle du comportement des utilisateurs. On suppose qu’il existe un "surfeur aléatoire" à qui une page web est attribuée au hasard. Ce surfeur clique sur des liens sans jamais revenir en arrière, mais finit par s’ennuyer et commence sur une autre page aléatoire. La probabilité que le surfeur aléatoire visite une page correspond à son PageRank.

Ce qu’ils impliquent, c’est qu’un surfeur aléatoire devrait visiter les pages importantes plus fréquemment et les pages non importantes moins souvent.
La manière d’interpréter cela précisément est la suivante :

1. Former le graphe qui encode les connexions entre les pages web renvoyées pour une requête particulière.
2. Construire une chaîne de Markov correspondant à une marche aléatoire sur ce graphe.
3. Construire une chaîne de Markov correspondant à une marche aléatoire sur ce graphe.

Essayons donc de mettre cela en pratique et de voir ce qui se passe.

a. La fausse évidence des liens entrants

Le surfeur aléatoire est au cœur de l’algorithme PageRank. Ce modèle simule le comportement d’un utilisateur fictif qui navigue sur le Web en alternant entre deux actions possibles :

- Cliquer sur un lien sortant : Avec une probabilité α (généralement proche de 1), il choisit un lien hypertexte disponible sur la page actuelle pour accéder à une autre page.

- Se téléporter : Avec une probabilité $1 - \alpha$, il quitte la page actuelle pour se déplacer aléatoirement vers une autre page.

Ce mécanisme de téléportation résout les problèmes liés aux sinks et aux sous-graphes isolés, tout en garantissant l'accessibilité de toutes les pages.

b. Google matrix

La Google Matrix est la représentation mathématique du comportement du surfeur aléatoire. Elle est construite à partir de deux éléments essentiels :

- La matrice de transition W : Cette matrice décrit les probabilités de navigation entre les pages via les liens hypertextes. Si une page j possède k liens sortants, la probabilité de passer de j à i est donnée par:

$$W_{ij} = \begin{cases} \frac{1}{k} & \text{si un lien existe entre la page } i \text{ et la page } j, \\ 0 & \text{sinon .} \end{cases}$$

- La matrice de téléportation: Cette matrice, notée R, représente la probabilité uniforme de sauts aléatoires vers n'importe quelle page. Chaque élément est défini par:

$$R_{ij} = \frac{1}{N}, \text{ où } N \text{ est le nombre total de pages.}$$

La Google Matrix M est alors définie comme une combinaison linéaire des deux matrices:

$$M = \alpha W + (1 - \alpha)R$$

c. Convergence de la méthode PageRank

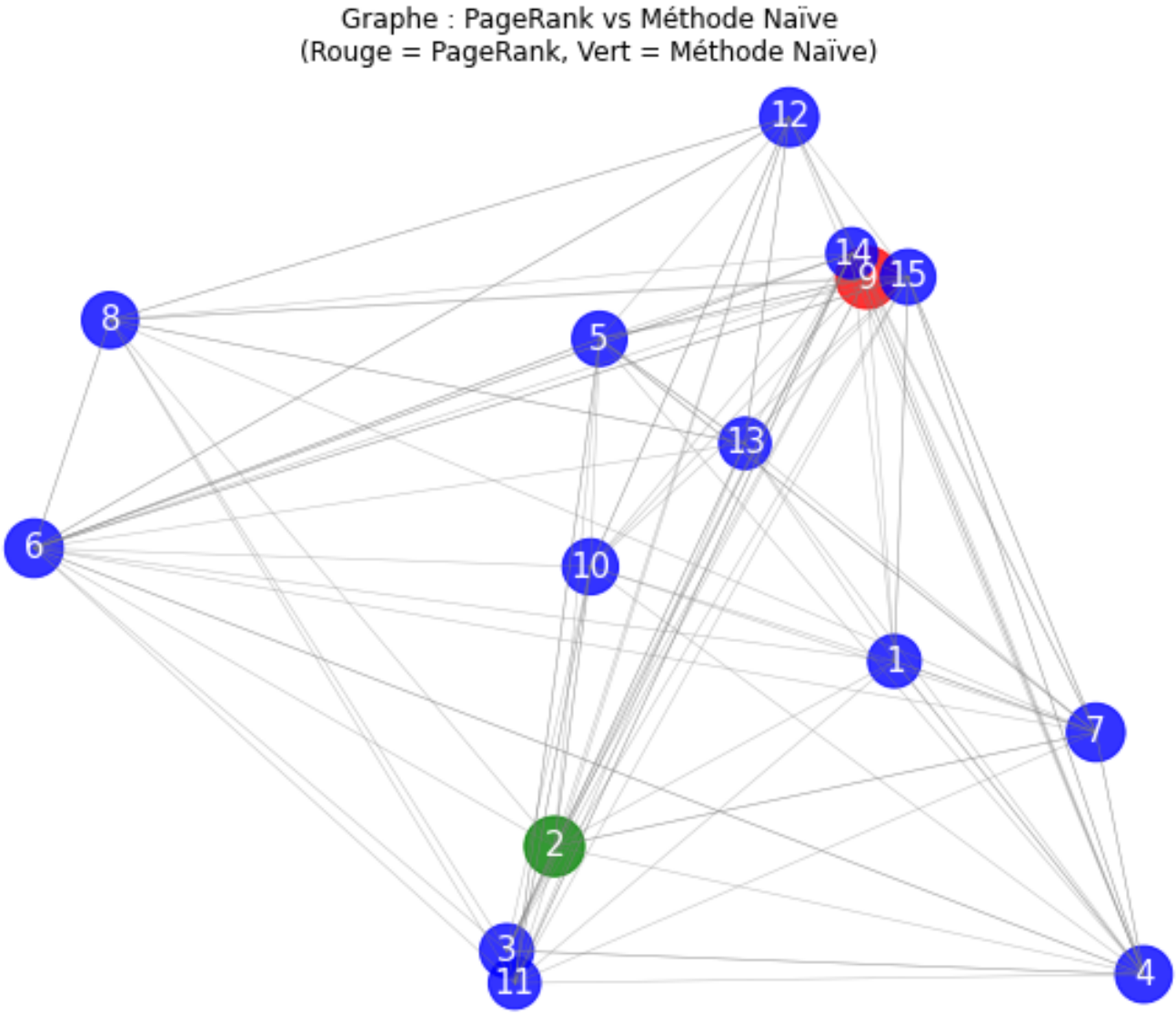
L'algorithme PageRank se base sur une chaîne de Markov ergodique, ce qui signifie qu'elle possède une unique distribution stationnaire. Concrètement, cela se traduit par l'existence d'un vecteur π qui vérifie :

$$\pi = \pi M, \text{ où } M \text{ désigne la Google Matrix.}$$

Pour trouver cette distribution, on part généralement d'une mesure de probabilité quelconque $\pi^{(0)}$, par exemple la loi uniforme dont les coordonnées valent $\frac{1}{N}$ (avec N le nombre de pages), puis on applique de manière itérative :

$$\pi^{(k+1)} = \pi^{(k)} M.$$

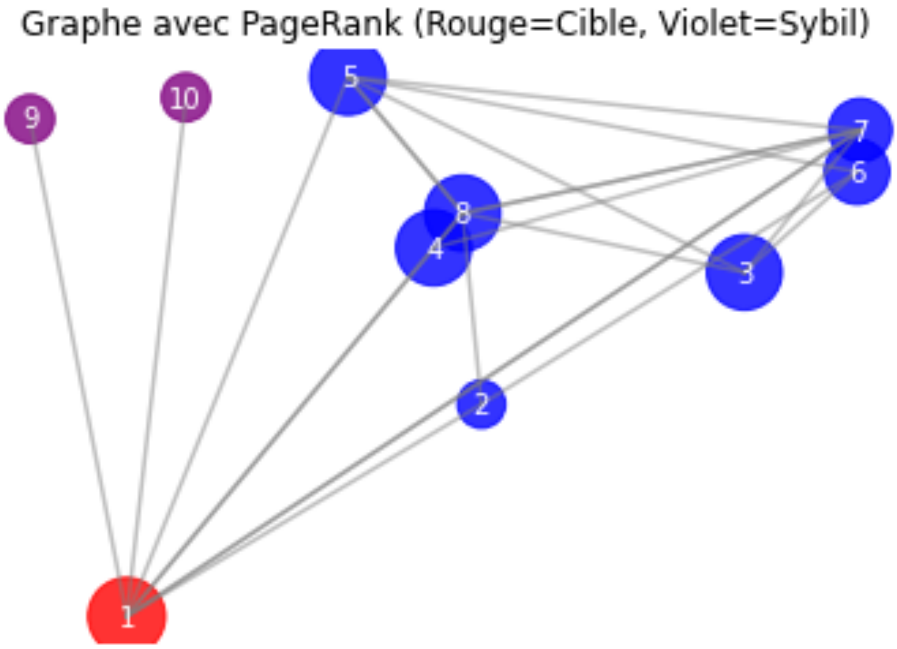
Pour déterminer le moment où la convergence est atteinte, on suit l'évolution de $\pi^{(k)}$ jusqu'à ce que la distance entre $\pi^{(k+1)}$ et $\pi^{(k)}$ devienne inférieure à un seuil ε. À ce stade, on considère que l'algorithme a convergé et que $\pi^{(k+1)}$ constitue la distribution PageRank finale, qui permet d'attribuer un score à chaque page.



PageRank a convergé en 9 itérations.

IV) Les attaques Sybil et le « Google Bombing »

Le principe d'une attaque Sybil est qu'une même entité se fait passer pour de multiples sites indépendants afin de fausser le PageRank. Concrètement, on crée des faux sites qui pointent tous vers une page cible, ou s'échangent des liens entre eux, dans le but de gonfler artificiellement son score. Ainsi, en augmentant le nombre de liens entrants, la page visée peut gagner en positionnement sans pour autant être réellement pertinente.



Le Google Bombing, quant à lui, repose davantage sur la manipulation des ancres de lien : on redirige massivement des liens dont le texte d'ancrage contient un mot-clé particulier, afin d'associer ce mot-clé à un site donné dans les moteurs de recherche. Cette pratique permet de fausser la manière dont les algorithmes interprètent la popularité ou la pertinence d'une page sur une requête donnée.

Les moteurs de recherche déploient donc des contre-mesures pour détecter ces manipulations et maintenir la qualité de leurs résultats.

V) Conclusion

L'algorithme PageRank a vraiment changé la façon dont nous organisons et classons les informations sur le web. Larry Page et Sergey Brin ont eu l'idée de s'inspirer de notre comportement en ligne pour créer un outil capable de mesurer l'importance d'une page. Contrairement aux méthodes de base qui se contentaient de compter le nombre de liens entrants, PageRank prend aussi en compte la qualité et la crédibilité des sites qui renvoient à une page. C'est ce qui en fait une méthode bien plus utile et pertinente pour nous.

Ce qui rend PageRank si intéressant, c'est son côté proche de nous : il s'appuie sur l'idée d'un « surfeur aléatoire » qui explore le web en cliquant sur des liens, mais qui, de temps en temps, décide de visiter une page au hasard. Ce mécanisme corrige les problèmes des pages « bloquées » ou isolées et garantit qu'on peut naviguer partout dans le réseau.

Évidemment, l'algorithme n'est pas parfait. Certains ont essayé de le manipuler avec des techniques comme les attaques Sybil ou le Google Bombing. Mais Google a su s'adapter et améliorer ses défenses pour continuer à nous offrir des résultats de recherche fiables et pertinents.

Aujourd'hui, même si PageRank n'est plus le seul critère utilisé pour classer les pages web, il reste un élément central. C'est grâce à des idées comme celle-là que nous pouvons trouver ce que nous cherchons dans l'immensité du web. En fin de compte, PageRank nous montre qu'une idée simple, inspirée de nos comportements, peut avoir un impact énorme sur notre quotidien en ligne.

VI) Bibliographie

- https://disco.ethz.ch/courses/ti2/lecture/markov_chains.pdf
- https://www.apmep.fr/IMG/pdf/AAA10065.pdf
- https://fr.wikipedia.org/wiki/PageRank
- https://lucidar.me/en/seo/pagerank-random-surfer/